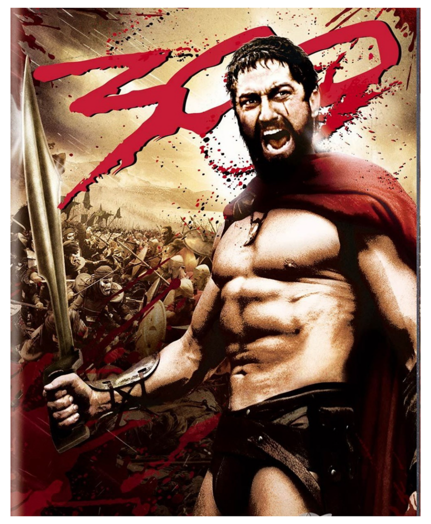

300-sparsans at SemEval-2018 Task 9: Hypernymy as interaction of sparse attributes

Gábor Berend[1], Márton Makrai[2], and Péter Földiák[3]

[1]Department of Informatics, University of Szeged berendg@inf.u-szeged.hu

[2]Research Institute for Linguistics of the Hungarian Academy of Sciences makrai.marton@nytud.mta.hu

[3]Secret Sauce Partners Peter.Foldiak@gmail.com

Sparse word representations

- motivation
  - focus on most salient parts of word representations (Faruqui et al., 2015; Berend, 2017; Subramanian et al., 2018)
  - increase separability, interpretability (Olshausen and Field, 1997) and stability against noise
- *Non-negative sparse coding*
  - for interpretability (Faruqui et al., 2015; Fyshe et al., 2015; Arora et al., 2016)

to describe the city of Pittsburgh, one might talk about phenomena typical of the city, like erratic weather and large bridges. It is redundant and inefficient to list negative properties, like the absence of the Statue of Liberty

(Subramanian et al., 2018)

  - in word translation (Berend, 2018)
    - sparse word vectors for the two languages such that coding bases correspond to each other

Formal concept analysis (FCA)

- FCA is the mathematization of a conceptual hierarchy
  - a set of *objects*, now words $w \in \mathcal{O}$,
  - a set of *attributes*, now word vector indices $i \in \mathcal{A}$, and
  - a binary incidence relation $\mathcal{I} \subseteq \mathcal{O} \times \mathcal{A}$, now $\langle w, i\rangle \in$ iff the ith coordinate in the sparse code of w is non-zero
- FCA finds formal *concepts*, pairs $\langle O, A\rangle$, $O \subseteq \mathcal{O}$, $A \subseteq \mathcal{A}$, such that
  - A consists of the shared attributes of objects in O (and no more), and
  - O consists of the objects in $\mathcal{O}$ that have all the attributes in A (and no more)
  - O and A are closed sets iff $\langle O, A\rangle$ is a concept
- O is called the extent and A is the intent of the concept
- order defined in the context: if $\langle O_i, A_i\rangle$ are concepts in C, $\langle O_1, A_1\rangle$ is a *subconcept* of $\langle O_2, A_2\rangle$ if $O_1 \subseteq O_2$ which is equivalent to $A_1 \supseteq A_2$
- lattice
- Adding attributes to $\mathcal{A}$, the original concepts will be embedded as a substructure
- The smallest node in the concept lattice $n(w)$ whose extent contains a word w is said to *introduce* the object
- h should be a hypernym of q iff $n(q) \leq n(h)$
- tools: Endres et al. (2010); Cimiano et al. (2005)
- features in the next column:
  - $n(w)$ is the concept that introduces w, i.e. the most specific location within the DAG for w
  - $n_1 \prec n_2$ denotes that n_1 is an immediate predecessor of n_2
  - Parents, and even the inverse relation, proved to be more predictive than the conceptually motivated $q \leq h$
  - not useful (see post-evaluation ablation experiments)

The task and our results

- extract hypernyms for query words Camacho-Collados et al. (2018)
- $(3\text{languages} + 2) \times 3$ subtasks
  - three languages, English, Italian, and Spanish
  - + two domains, medical and music
  - queries types: concepts, entities, or all
- Our system took first place in subtasks
  - *(1B) Italian (all* and *entities)*
  - *(1C) Spanish entities* and
  - *(2B) music entities*

Sparse vectors

- for each subtask, we solve for

$$\min_{D\in\mathcal{C},\alpha\in\mathbb{R}^{k\times|V|}_{\geq 0}} \|D\alpha - W_x\|_F + \lambda\|\alpha\|_1,$$

- $\mathcal{C}$ is the convex set of $\mathbb{R}^{d\times k}$ matrices with column norms ≤ 1, and
- α contains the sparse coefficients for the words
- akin to Berend (2017) + new non-negativity constraint over the elements of α
- To keep the size of the FCA tree manageable, we only included the query words and the training hypernyms. This restriction turns out to be very useful.
- dense embedding W unit-normed, $\lambda = .3$

Features summarized

for query q and its hypernym candidate h

dense vectors W_x	skip-gram in $d = 100$-dimensions
cosine	$\frac{\mathbf{q}^\intercal\mathbf{h}}{\|\mathbf{q}\|_2\|\mathbf{h}\|_2}$
difference	$\|\mathbf{q} - \mathbf{h}\|_2$
normRatio	$\frac{\|\mathbf{q}\|_2}{\|\mathbf{h}\|_2}$
word strings	
qureyBeginsWith	$Q[0] = h$
queryEndsWith	$Q[-1] = h$
hasCommonWord	$Q \cap H \neq \emptyset$
sameFirstWord	$Q[0] = H[0]$
sameLastWord	$Q[-1] = H[-1]$
logFrequencyRatio	$\log_{10}\frac{count(q)}{count(h)}$
isFrequentHypernym[1]	$c \in MF_{50}(q.type)$
FCA	see previous column
sameConcept	$n(h) = n(q)$
parent	$n(q) \prec n(h)$
child	$n(h) \prec n(q)$
sparse vectors	$\phi(w)$: set of non-zero coordinates, $k = 200$
overlappingBasis	$\phi(q) \cap \phi(h) \neq \emptyset$
sparseDifference$_{q\setminus h}$	$\|\phi(q) - \phi(h)\|$
sparseDifference$_{h\setminus q}$	$\|\phi(h) - \phi(q)\|$
attributePair$_{ij}$	$\langle i, j\rangle \in \phi(q) \times \phi(h)$

- $MF_{50}(q.type)$: 50 most frequent hypernyms for the query type (i.e. concept or entity). Debugged after submission.
- `attributePair`$_{ij}$s are the most important features
  - indicator features for the interaction terms between the sparse coefficients in α
  - This feature template induces k^2 features, with k being the number of basis vectors introduced in the dictionary matrix D according to Eq. 1
  - the role of these features is similar to *interaction terms* in regression

Two submissions

- one of our submissions involved attribute pairs, the other not
- both submissions used the FCA-based features
  - conceptually motivated but practically harmful

Implementation and tricks

- dense vectors: skip-gram (Mikolov et al. 2013, $d = 100$) trained for each sub-corpus provided by the organizers
- multi-token phrases with the `word2phrase` software accompanying w2v
- top 15 selected by logistic regression trained for concepts and entities
  - `sklearn` (Pedregosa et al., 2011),regularization parameter set to the default 1.0
- For each training pair (q, h), we generated a number of negative samples (i.e. the training data does not include h' as a valid hypernym for q)
  - h' sampled from the valid training hypernyms in the query type (*concept* or *entity*)
- post-ranking heuristic
  - re-ranking according to background frequency in the training corpus
  - motivation: more frequent words refer to more general concepts and more general hypernymy relations may be more easily to detect
- OOV backoff by query type

Post-evaluation analysis (without FCA)

Features derived with sparse attribute pairs and/or FCA:

		MAP	MRR	P@1	P@3	P@5	P@15
off	off	10.3	21.3	15.0	10.6	10.1	9.6
off	on	10.1	21.1	14.9	10.5	9.9	9.5
on	off	**12.1**	25.4	**18.9**	12.9	**11.6**	10.9
on	on	**12.1**	**25.3**	18.7	**13.0**	**11.6**	**11.0**

- number of basis vectors in sparse coding ($k \in \{\mathbf{200}, 300, 1000\}$),
- number of negative training samples per positive sample
  - submissions: **50** negative samples generated per query q
  - post evaluation: all hypernyms in the training set except for the proper hypernyms for q
- candidates filtered to those present in the training data
  - Historically, applied to speed up the FCA algorithm (smaller concept lattice)
- **boldface font** above: submission settings

1A

		candidate filtering off						candidate filtering on					
k	ns	MAP	MRR	P@1	P@3	P@5	P@15	MAP	MRR	P@1	P@3	P@5	P@15
200	50	6.5	14.9	13.1	7.4	6.1	5.5	12.1	25.4	18.9	12.9	11.6	10.9
200	all	6.9	15.8	14.1	7.6	6.3	5.8	13.0	27.1	19.9	14.2	12.5	11.8
300	50	6.9	15.8	13.9	7.6	6.4	5.9	12.1	25.7	19.5	13.0	11.5	11.0
300	all	8.0	17.8	15.4	8.9	7.4	6.8	13.5	28.0	21.1	14.5	12.9	12.3
1000	50	9.0	20.0	17.2	9.8	8.3	7.7	13.3	**28.1**	**21.3**	13.8	12.6	12.3
1000	all	11.6	26.1	22.5	12.5	10.8	10.0	**13.6**	27.2	19.4	**13.9**	**13.2**	**12.8**

`https://github.com/begab/fca_hypernymy`

Research supported by the project *Integrated program for training new generation of scientists in the fields of computer science*, no. EFOP-3.6.3-VEKOP-16-2017-0002. The project has been supported by the European Union and co-funded by the European Social Fund.

(Post-evaluation conted:) All the subtasks

	MAP	MRR	P@1	P@3	P@10	P@15
1A	13.3	28.1	21.3	13.8	12.6	12.3
1A	19.8	36.1	29.7	21.1	19.0	18.3
1B	**12.5**	24.2	14.5	**13.4**	**12.5**	**12.0**
1B	12.1	**25.1**	**17.6**	12.9	11.7	11.2
1C	**21.8**	**43.8**	**33.7**	**22.9**	**21.4**	**19.9**
1C	20.0	28.3	21.4	20.9	21.0	19.4
2A	21.9	39.5	34.2	25.5	22.6	18.5
2A	34.0	54.6	49.2	40.1	36.8	27.1
2B	31.5	43.6	29.8	30.3	30.3	31.5
2B	41.0	60.9	48.2	44.9	41.3	38.0

- upper: our system, ($k = 1000, ns = 50$, hypernym candidate filtering on, FCA off)
- lower: subtask winner, official scores

Future work: hierarchical sparse coding

- trees describe the order in which variables "enter the model" (i.e., take non-zero values, Zhao et al. (2009))
- a node may enter only if its ancestors also do
- top level nodes should focus on *general* meaning components
- efficient implementation (Yogatama et al., 2015)
- correspondence between the variable tree and the hypernym hierarchy

References

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Linear algebraic structure of word senses, with applications to polysemy. *arXiv:1601.03764v1*, 2016.

G. Berend. Sparse coding of neural word embeddings for multilingual sequence labeling. *Transactions of the Association for Computational Linguistics*, 5:247–261, 2017. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1063.

G. Berend. Towards cross-lingual utilization of sparse word representations. In V. Vincze, editor, *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*, pages 272–280. Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2018.

J. Camacho-Collados, C. Delli Bovi, L. Espinosa-Anke, S. Oramas, T. Pasini, E. Santus, V. Shwartz, R. Navigli, and H. Saggion. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States, 2018. Association for Computational Linguistics.

P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal Artificial Intelligence Research (JAIR)*, 24:305–339, 7 2005.

D. Endres, P. Földiák, and U. Priss. An Application of Formal Concept Analysis to Semantic Neural Decoding. *Annals of Mathematics and Artificial Intelligence*, 57(3-4):233–248, 07 2010. doi: 10.1007/s10472-010-9196-8. reviewed.

M. Faruqui, J. Dodge, S. Jauhar, C. Dyer, E. Hovy, and N. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL 2015*, 2015. Best Student Paper Award.

A. Fyshe, L. Wehbe, P. P. Talukdar, B. Murphy, and T. M. Mitchell. A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–41, 2015.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, 05 2013. URL http://arxiv.org/abs/1301.3781.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, and E. Hovy. Spine: Sparse interpretable neural embeddings. *AAAI*, 2018.

D. Yogatama, M. Faruqui, C. Dyer, and N. A. Smith. Learning word representations with hierarchical sparse coding. In *ICML*, 2015. Previous version in NIPS Deep Learning and Representation Learning Workshop 2014.

P. Zhao, G. Rocha, and B. Yu. The composite and absolute penalties for grouped and hierar-